Methods for Evaluating Inter-Rater Agreement During the NLM/AHCPR Large Scale Vocabulary Test

Eduardo Ortiz, M.D., M.P.H., J. Chris Eagon, M.D. M.S., Michael J. Lincoln, M.D.
Department of Medical Informatics: VA Medical Center, VA IRM Field Office, and University of Utah Health Sciences Center, Salt Lake City, Utah

Introduction. The Large Scale Vocabulary Test (LSVT) was developed by the National Library of Medicine to determine the extent to which clinically relevant vocabulary is covered by the UMLS Metathesaurus and Planned Additions [1]. A consortium of researchers was organized at the Department of Veterans Affairs and University of Utah to participate in the LSVT. Because of the large numbers of raters involved, it was important to maintain consistency among raters in the analysis of terms. The aim of this study was to develop a method for establishing and maintaining inter-rater agreement in order to ensure the validity of our results on the LSVT.

Methods. Seventeen raters participated in the study. A set of explicit rules and coded comments was developed to help raters maintain consistency in their ratings and allow them to characterize the differences between the submitted term and the LSVT match term. A series of test sets was also developed to train the raters in the use of the LSVT interface and on the appropriate use of rules and comments. Test set terms were selected to illustrate the various interface responses (exact match vs approximate match vs no match) and to provide raters with varying degrees of matching difficulty (easy vs difficult).

Raters completed Test Set I the first week of the study in interactive sessions with a preceptor. Raters then independently rated Test Set II. After completing TS II, all seventeen raters met in group sessions to discuss the LSVT responses for each of the submitted terms, resulting in some modifications to the rules and comments. Each of the raters then independently rated 50 to 200 of the data set terms before going on to complete Test Set III. No modifications to the rules or comments were made after completion of TS III. All of the raters' responses were collected and loaded into Microsoft Access and Excel. Inter-rater agreement was analyzed using the modal response rate for each decision made by a rater compared to the mean modal response rate of the group. Effects on modal response rates due to match type, match difficulty, and test set were determined using repeated measures analysis of variance (ANOVA).

Results. Mean modal response rates ranged from 63% for the most difficult terms to 99% for the easiest terms. Overall, the mean modal response rate for all 38 decisions made by a rater was 81% in TS II and 82% in TS III. For both test sets, easy terms resulted in a higher mean modal response rate than difficult terms ($F(1,16) = 48$; $p < .0001$). Exact matches also resulted in a higher rate than approximate matches ($F(1,16) = 29$; $p < .0001$). No significant differences in mean modal response rates were found between test sets II and III ($F(1,16) = 1$; $p < .337$). There were also no other significant interactions between the test sets in terms of match type and term difficulty.

Conclusion. An assessment of inter-rater agreement is required when doing rating work as in the LSVT. A large part of our initial work was therefore focused on these efforts. All seventeen raters achieved a level of agreement within two standard deviations of the group mean on test set terms. Our results show that raters can be trained to evaluate terms in a consistent manner. We believe that our results support the conclusion that inter-rater agreement was acceptably good for the LSVT work we performed.

References

1) Humphries BL, Hole WT, McCray AT, Fitzmaurice JM. Planned NLM/AHCPR Large-Scale Vocabulary Test: Using UMLS Technology to Determine the Extent to Which Controlled Vocabularies Cover Terminology Needed for Health Care and Public Health. JAMIA. 1996; 3: 281-87.

1091—8280/97/$5.00